# Literature Review

Dozens of algorithms and techniques have been proposed to address imbalanced data classification problems. Some of the algorithms are based on preprocessing the data itself in order to achieve balance between the classes, whereas others are based on modifying the machine learning algorithms to pay more attention to the minority class. Both ways aim to make the models more fair and accurate.

This chapter will thoroughly review these techniques and analyze their impact on the performance of machine learning models. We will analyze various studies, assessing their methodologies, the datasets they used, and their findings. Through this approach, our goal is to gain a deeper understanding of the advancements achieved, the aspects that are working well, and the opportunities for enhancement.

Haixiang et al. present a thorough examination of methods and applications for dealing with unbalanced data in their study titled "Learning from class imbalanced data: Review of methods and applications". The study discusses various strategies such as preprocessing and cost-sensitive learning to handle imbalanced datasets. It also delves into classification models, including ensemble classifiers and algorithmic modified classifiers. The paper highlights the importance of multi-class classification in extending binary classification algorithms. The research methodology involved a systematic search process to compile relevant papers, leading to a detailed analysis of trends in imbalanced learning over the past decade. The study covers a wide range of application domains, from financial management to medical diagnosis and telecommunications, showcasing the diverse practical implications of addressing class imbalance in data analysis.

In their paper titled "SMOTE: Synthetic Minority Over-sampling Technique", Chawla, Bowyer, Hall, and Kegelmeyer address the challenge of imbalanced datasets in classification tasks and propose a new method to deal with the class imbalance. The SMOTE method combines under-sampling of the majority class with a specialized form of over-sampling the minority class to improve classifier performance.

They compare their approach with other re-sampling techniques and demonstrate its effectiveness using various datasets and classifiers. The study highlights the importance of addressing class imbalance in machine learning and provides insights into potential solutions for improving classification accuracy in real-world applications.

The paper "Adaptive Synthetic Sampling Approach for Imbalanced Learning" by He et al. introduces the ADASYN algorithm, a novel method for addressing imbalanced data classification problems. ADASYN dynamically generates synthetic data samples for the minority class to reduce bias and improve learning performance. ADASYN addresses the challenge of learning from imbalanced data sets by using a weighted distribution approach for minority class examples based on their level of difficulty in learning. It generates more synthetic data for minority class examples that are harder to learn compared to those that are easier to learn. This adaptive synthetic sampling approach aims to reduce bias introduced by class imbalance and shift the classification decision boundary towards difficult examples, thereby improving learning performance. The simulation results and assessment metrics both indicate the algorithm's performance on the real-world data sets from the UCI Machine Learning Repository. The paper highlights the significance of unbalanced learning in machine learning research and proposes that ADASYN could serve as a powerful tool for studying numerous classes and incremental imbalanced learning.

In their research paper titled "Boosting Support Vector Machines for Imbalanced Data Sets," Wang and Japkowicz describe a novel approach that makes use of support vector machines and boosting algorithms to improve the accuracy of predictions made for both majority and minority classes. In order to avoid the limitations of under- and oversampling, they place an emphasis on the distribution of data and the change of classifier modifications. The study demonstrates that their ensemble classifier approach is capable of solving skewed vector spaces and overfitting in an effective manner, which has the promise of providing a solution for imbalanced data sets in applications that are used in the real world.

The research paper titled "Balancing Training Data for Automated Annotation of Keywords: a Case Study" by Batista, Bazzan, and Monard discusses the process of learning from imbalanced data sets in SWISS-PROT keyword annotation. The authors highlight the difficulties faced when dealing with skewed class distributions. They discuss the impact of imbalanced data on machine learning algorithms such as k-Nearest Neighbor and decision trees, emphasizing the need for data balancing techniques to improve model performance.

To address the issue of class imbalance, the research suggests using random oversampling as well as SMOTE oversampling combined with Tomek Links. The findings indicate that there has been a reduction in the number of false negatives and a slight increase in the number of false positives.

Dal Pozzolo, Caelen, and Bontempi (2015) provides a comprehensive analysis of the effectiveness of undersampling as a solution for class imbalance in their paper titled "When is Undersampling Effective in Unbalanced Classification Tasks?". The study delves into the impact of undersampling on classification accuracy by considering the increase in variance due to sample reduction and the warping of the posterior distribution caused by changes in prior probabilities. The findings suggest that the effectiveness of undersampling is influenced by factors such as the number of samples, classifier variance, class imbalance degree, and posterior probability value. The research highlights the need for a nuanced approach to undersampling, recommending adaptive selection techniques like racing for tailored and calibrated undersampling strategies. This paper contributes valuable insights to the exploration of class imbalance solutions, emphasizing the importance of considering various factors and adopting customized approaches to enhance classification effectiveness in unbalanced datasets.

The authors recommend a cautious and tailored approach to rebalancing classes before learning a classifier in unbalanced classification scenarios. Specifically, they suggest the following strategies:

1. Avoid Naive Undersampling: The authors caution against a simplistic or unsupervised application of undersampling, as it may not always lead to improved classification performance. Instead, they advocate for a more informed and adaptive selection of undersampling techniques
2. Adopt Specific Adaptive Selection Techniques: The authors suggest using specific adaptive selection strategies, such racing, to successfully alleviate class imbalance. With the help of these methods, undersampling strategies can be evaluated and calibrated on an individual basis depending on the needs of the learning job and the dataset's properties.
3. Consider Prior Probabilities and Variances: The authors talk about how crucial it is to take into account how undersampling may affect the dataset's prior probabilities and variances. Making decisions about rebalancing strategies and their possible impact on classification accuracy can be aided by having a thorough understanding of these effects.

By recommending a thoughtful and adaptive approach to rebalancing classes in unbalanced classification scenarios, the authors aim to enhance the effectiveness of undersampling techniques and improve the overall performance of classifiers trained on imbalanced datasets.

The paper "Ranked Minority Oversampling in Boosting (RAMOBoost)" by Chen et al. introduces a novel approach to address class imbalance in machine learning. RAMOBoost combines oversampling techniques with boosting to generate synthetic instances based on the class ratios of nearest neighbors in the minority class. Unlike traditional methods, RAMOBoost adaptively adjusts sampling weights and focuses on difficult-to-learn instances in both majority and minority classes. The study highlights the importance of handling both relative and absolute imbalances in imbalanced learning problems. While the evaluation is limited to two-class imbalanced scenarios, the authors suggest potential extensions to handle multiclass imbalances. The use of Euclidean distance as the distance measure is noted, with room for exploring alternative measures. The paper emphasizes the need for parameter tuning and suggests ongoing research to address these challenges. Overall, RAMOBoost shows promise as a powerful tool for imbalanced learning, offering new insights and potential applications across various domains.

The paper by Krawczyk provides a comprehensive overview of the challenges and future directions in learning from imbalanced data. It discusses the evolution of imbalanced learning beyond skewed binary tasks, highlighting the importance of understanding the nature of imbalanced data in contemporary real-world applications. The author emphasizes the need for improved data-level and algorithm-level methods, as well as hybrid approaches, to address the complexities of imbalanced datasets. The paper identifies seven vital areas of research, including classification, regression, clustering, data streams, and big data analytics, offering insights into open challenges and potential solutions. Overall, the paper calls for a deeper analysis of minority class structures, development of multi-class and multi-label learning methods, efficient clustering techniques, and enhanced evaluation measures for imbalanced regression problems.

The paper "Learning from Imbalanced Data" by He & Garcia provides a comprehensive review of the challenges and advancements in addressing the imbalanced learning problem. It discusses the critical nature of learning from imbalanced data in various real-world applications such as surveillance, security, Internet, and finance. The paper highlights the need for new understandings, principles, algorithms, and tools to effectively handle underrepresented data and severe class distribution skews. It covers the state-of-the-art technologies and assessment metrics used to evaluate learning performance in imbalanced scenarios.

In their publication titled "SMOTEBoost: Improving Prediction of the Minority Class in Boosting", Chawla et al. present a novel approach for tackling class imbalance in datasets by integrating the SMOTE algorithm with boosting approaches. SMOTEBoost utilizes the technique to generate synthetic samples from the minority class using SMOTE after each boosting round in order to rebalance the distribution and enhance prediction performance. Empirical findings on diverse datasets indicate that SMOTEBoost surpasses conventional boosting algorithms and standalone SMOTE applications, attaining superior F-values and improving prediction accuracy for the minority class. This approach is especially advantageous for domains that involve rare occurrences, such as network intrusion detection and medical diagnostics, where precise identification of minority classes is essential. The work emphasizes the significance of taking into account class imbalance in machine learning tasks and offers a feasible strategy for improving predictive performance on minority classes.

Brown and Mues introduces two novel algorithms in he paper "Exploratory Undersampling for Class-Imbalance Learning", EasyEnsemble and BalanceCascade, designed to address the limitations of traditional undersampling methods in handling imbalanced data sets. These algorithms have demonstrated superior performance in terms of Area Under the ROC Curve, F-measure, and G-mean values compared to existing class-imbalance learning methods. Moreover, they offer faster training times, making them efficient solutions for dealing with imbalanced data sets.

The paper "RUSBoost: Improving Classification Performance when Training Data is Skewed" by Seiffert et al., introduces RUSBoost, a novel algorithm that combines random undersampling with boosting to enhance classification performance in the presence of skewed training data. The study compares RUSBoost to SMOTEBoost, highlighting RUSBoost's simplicity, computational efficiency, and faster model training times. Results demonstrate that RUSBoost outperforms SMOTEBoost on most datasets, offering a promising alternative for addressing class imbalance. Future research directions include evaluating RUSBoost with additional learners and datasets, as well as comparing its performance with other boosting algorithms tailored for class imbalance.

The paper "Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction" by Weiss and Provost contributes to the exploration of class imbalance solutions by investigating the impact of class distribution on classifier performance in scenarios where training data are limited and costly to obtain. The study examines the relationship between class distribution, accuracy, and AUC, providing insights into the effectiveness of data balancing techniques on model performance. The authors find that the optimal class distribution for accuracy tends to align with the natural class distribution, while for AUC, it tends to be closer to a balanced distribution. They introduce a budget-sensitive progressive sampling algorithm to enhance classifier performance when data procurement and learning can be integrated. The research underscores the importance of selecting suitable class distributions in training data to enhance classifier performance, particularly in situations with limited training examples and expensive data acquisition. The findings suggest that as the volume of training data increases, the performance differences across various class distributions diminish, highlighting the decreasing significance of class distribution as more data becomes accessible. Overall, the study offers valuable insights and guidelines for leveraging class imbalance solutions to improve model performance in contexts where data balancing techniques are crucial for addressing class imbalance challenges.

As we discussed in the first chapter, this research also aims to test anomaly detection algorithms for tackling the class imbalance problem. Following reviewed works are related to anomaly detection algorithms which we used in this research as well.

Breunig et al. introduce a novel method for detecting outliers in multidimensional datasets by introducing the concept of local outlier factor (LOF) in their paper "Identifying Density-Based Local Outliers". Unlike traditional outlier detection methods that treat outliers as binary properties, LOF quantifies the degree of outlyingness for each object in the dataset. The approach is based on the local density of objects in their neighborhoods, without the need for explicit cluster information. The paper provides a thorough analysis of the formal properties of LOF, including bounds on outlier factors and guidelines for choosing the parameter MinPts. Experimental results demonstrate the effectiveness and efficiency of LOF in identifying local outliers that may be overlooked by other methods. The study highlights the importance of considering outliers with varying degrees of outlyingness in data analysis applications.

The paper "Isolation Forest" by Liu et al., introduces a novel model-based anomaly detection method called Isolation Forest (iForest), which focuses on isolating anomalies rather than profiling normal instances. By exploiting the characteristics of anomalies being 'few and different', iForest efficiently builds an ensemble of Isolation Trees (iTrees) to detect anomalies with short average path lengths. This approach offers advantages such as linear time complexity, low memory requirement, and scalability to handle large data sets and high-dimensional problems. Empirical evaluations demonstrate iForest's favorable performance compared to existing methods in terms of Area Under the Curve (AUC) and processing time. The paper provides insights into the efficiency and effectiveness of iForest in detecting anomalies, making it a promising approach for anomaly detection tasks.

**References**

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, 220-239.

Chawla, Nitesh & Bowyer, Kevin & Hall, Lawrence & Kegelmeyer, W.. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. (JAIR). 16. 321-357. 10.1613/jair.953.

Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Hong Kong, 2008, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.

keywords: {Classification algorithms;Decision trees;Algorithm design and analysis;Training data;Machine learning;Accuracy;Machine learning algorithms},

Wang, B. X., & Japkowicz, N. (2009, March 5). Boosting support vector machines for imbalanced data sets. *Knowledge and Information Systems*, *25*(1), 1–20. <https://doi.org/10.1007/s10115-009-0198-y>

Batista, Gustavo & Bazzan, Ana & Monard, Maria-Carolina. (2003). Balancing Training Data for Automated Annotation of Keywords: a Case Study.. the Proc. Of Workshop on Bioinformatics. 10-18.

Dal Pozzolo, Andrea & Caelen, Olivier & Bontempi, Gianluca. (2015). When is Undersampling Effective in Unbalanced Classification Tasks?. 10.1007/978-3-319-23528-8\_13.

S. Chen, H. He and E. A. Garcia, "RAMOBoost: Ranked Minority Oversampling in Boosting," in IEEE Transactions on Neural Networks, vol. 21, no. 10, pp. 1624-1642, Oct. 2010, doi: 10.1109/TNN.2010.2066988.

keywords: {Nearest neighbor searches;Boosting;Euclidean distance;Training;Distribution functions;Clustering algorithms;Adaptive boosting;data mining;ensemble learning;imbalanced data},

Krawczyk, B. Learning from imbalanced data: open challenges and future directions. *Prog Artif Intell* **5**, 221–232 (2016). <https://doi.org/10.1007/s13748-016-0094-0>

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

Chawla, Nitesh & Lazarevic, Aleksandar & Hall, Lawrence & Bowyer, Kevin. (2003). SMOTEBoost: Improving Prediction of the Minority Class in Boosting. Proceedings of the 7th European conference on principles and practice of knowledge discovery in database. 2838. 107-119. 10.1007/978-3-540-39804-2\_12.

X. -Y. Liu, J. Wu and Z. -H. Zhou, "Exploratory Undersampling for Class-Imbalance Learning," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 39, no. 2, pp. 539-550, April 2009, doi: 10.1109/TSMCB.2008.2007853.

keywords: {Machine learning;Learning systems;Error analysis;Costs;Data mining;Research and development;Laboratories;Credit cards;Sampling methods;Class-imbalance learning;data mining;ensemble learning;machine learning;undersampling;Class-imbalance learning;data mining;ensemble learning;machine learning;undersampling},

C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse and A. Napolitano, "RUSBoost: Improving classification performance when training data is skewed," 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 2008, pp. 1-4, doi: 10.1109/ICPR.2008.4761297. keywords: {Training data;Boosting;Sampling methods;Data mining;Diseases;Costs;Iterative algorithms;Voting;Algorithm design and analysis},

Weiss, G. M., F., Provost. "Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction". *Journal of Artificial Intelligence Research* 19. (2003): 315–354.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May 16). LOF. ACM SIGMOD Record, 29(2), 93–104. https://doi.org/10.1145/335191.335388

Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. “Isolation forest.” Data Mining, 2008. ICDM’08. Eighth IEEE International Conference on.